John von Neumann Institute for Computing

Forcefield Validation with the Rosetta Protein Decoy Set

A. Verma, W. Wenzel

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 283-286, 2007.

http://www.fz-juelich.de/nic-series/volume36

Forcefield Validation with the Rosetta Protein Decoy Set

Abhinav Verma and Wolfgang Wenzel

Institute for Nanotechnology,
Research Center Karlsruhe, 76344 Karlsruhe, Germany
E-mail: {*verma, wenzel*}*@int.fzk.de*

We have recently extended our helical protein forcefield PFF01 to a more generalized protein forcefield PFF02 in our efforts towards a universal free energy forcefield for all atom protein folding and prediction. Here we selectivity of various proteins PFF02 with a Rosetta decoy set consisting of 32 proteins. The results conclude good selectivity of PFF02 for structure prediction with an average z-score of -3.46 and an average root mean square deviation of 2.14 Å.

1 Introduction

All atom protein folding and structure prediction have been one of the central problems in biophysical chemistry. Transferable potentials are needed to address these questions for a wide range of proteins[1]. We have recently extended our helical protein forcefield PFF01[2] to a more generalized protein forcefield PFF02[3] following the thermodynamic hypothesis[4], that most proteins are in thermodynamic equilibrium with the environment. With PFF02, we could demonstrate folding of small hairpin polypeptides into their native-like conformations[5,6].

The accuracy and predictivity of free energy protein forcefields can be investigated using decoy sets[7], a method that works even for proteins that are too large or too complex to be folded from random initial conformations. For the selectivity of PFF02, we study a decoy set generated using Rosetta[8] consisting of 32 proteins.

2 Method

A decoy set is a large library of protein conformations generated to approximately span all relevant low energy regions of the free energy landscape. To measure the predictivity and selectivity of a forcefield, the conformations in the library (decoy set) must be ranked according to their energy. If near native conformations emerge lowest in the free-energy function, the force field differentiates between native and near-native conformations. In the limit of completeness of the decoy set, which is rarely reached in practice, this test alone is sufficient to show that the force field stabilizes the native conformation of the protein against all competing metastable conformations and corresponds to the global optimum of the free-energy force field.

For decoy sets generated with unbiased methods, the computation of the Z-score (the difference between energies of near-native decoys to the mean energy of the decoy set in units of its standard deviation) gives a quantitative measure of the selectivity of the force field. The Z-score is defined as

$$Z = \frac{E_{\text{ref}} - \langle E \rangle}{\sigma} \tag{1}$$

where E_{ref} is the reference energy, *i.e.*, the energy of the native conformation, $\langle E \rangle$ is the average energy of the decoy set and σ is the standard deviation of the decoy set. The Z-score simply measures the mean energy distance from the native state of protein in terms of the standard deviations of the decoy set. The lower the Z-score, the better is the discrimination between native and non-native conformations in the decoy set. The histograms showing the distribution of decoys over energy range are shown in Figure 1(inset).

3 Results

For this study we investigated, which of the proteins of the the large all atom Rosetta decoy sets[9] could be stabilized by PFF02. The proteins in this decoy set range between 32-85 amino acids in size and span all secondary structural classes.

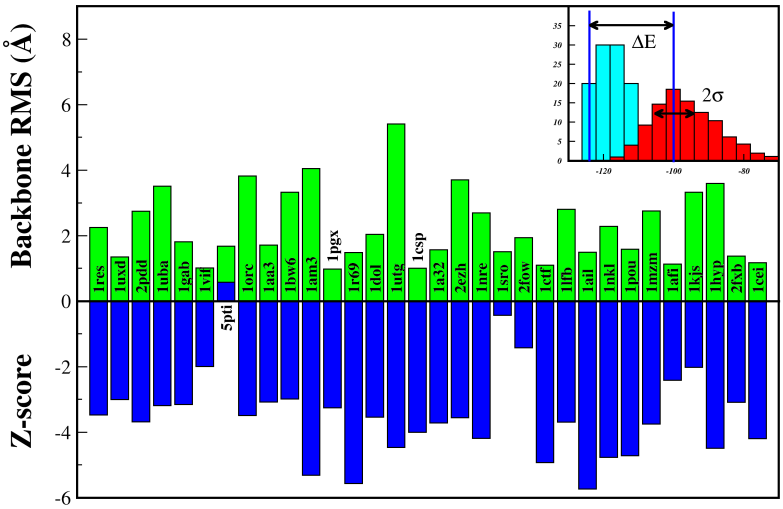

Figure 1. RMSD of the lowest energy conformation (Green) and Z-scores (Blue) of proteins in the Rosetta decoy set. Inset shows a sample distribution of decoys for a protein. The bars in cyan represent the distribution of near-native decoys generated from native structure and red bars represent all the decoys from the decoy set.

For the calculation of Z-scores we generated near-native conformations for 32 proteins of the latest Rosetta decoy library. We excluded only proteins that are stabilized by transition metal clusters or other ligands as such interactions are yet to be implemented in the present force field. The resulting near-native conformations deviate 1-4 Å from the experimental conformation, except for 1am3 and 1utg, where deviations of 4.05 and 5.4 Å

respectively are observed (top panel of Figure 1, Table 1 for all data). Since both of these proteins are dimeric, this difference arises because the molecules are relaxed here in isolation. The average deviation between experiment and near-native conformation in the force field for the set of 32 proteins was 2.14 Å, the figure also indicates that there is little correlation between the size of the protein and the accuracy with which the local minimum of the force field agrees with the experimental conformation.

In order to arrive at a meaningful comparison of the energies we relaxed the approximately 2000 decoys for each of the proteins in the decoy library in PFF02. This procedure maps each decoy to a local minimum of the force field of similar structure, the average change in RMSD between the starting and relaxed conformation was less than 0.02 Å. This means that the decoys are not changed in the relaxation process.

PDB ID	Z-Score	RMSD (Å)	PDB ID	Z-Score	RMSD (Å)
1a32	-3.72	1.57	1nre	-4.19	2.69
1aa3	-3.08	1.71	1orc	-3.49	3.82
1afi	-2.41	1.13	1pgx	-3.26	0.98
1ail	-5.73	1.49	1pou	-4.72	1.58
1am3	-5.32	4.05	1r69	-5.57	1.48
1bw6	-2.98	3.32	1res	-3.47	2.25
1cei	-4.19	1.17	1sro	-0.43	1.51
1csp	-4.01	1.00	1uba	-3.19	3.96
1ctf	-4.93	1.10	1utg	-4.47	5.41
1dol	-3.54	2.04	1uxd	-3.00	1.35
1gab	-3.16	1.81	1vif	-2.00	1.01
1hyp	-4.49	3.59	2ezh	-3.56	3.70
1kjs	-2.02	3.32	2fow	-1.43	1.94
1lfb	-3.69	2.80	2fxb	-3.09	1.37
1mzm	-3.75	2.75	2pdd	-3.69	2.74
1nkl	-4.77	2.28	5pti	0.58	1.68

Table 1. Zscores and RMSD(lowest energy) for the 32 proteins of Rosetta decoy set in PFF02.

The Z-scores for 29 out of the 32 proteins in the decoy set are less than -2.0 (top panel of Figure 1). This indicates a good selectivity of the force field for these proteins. The average the score of -3.46 is lower than that of any previously reported alternate scoring function for the same decoy set. The average Z-score for the same set of proteins in PFF01 was -3.06[10]. This indicates the improvement of the force field for this set of proteins which spans all kinds of secondary structural elements, with the only exception of 5PTI. Since the Rosetta decoy sets were specifically generated to span a wide range of near-native and non-native conformations for each protein. These data indicate that PFF02 stabilizes near-native conformations of a large family of small and medium-size proteins of all secondary structure classes as its global optimum.

4 Concluding Remarks

In this work, a 32 protein Rosetta decoy set was used to test the selectivity and predictivity of a recently modified protein forcefield PFF02. The results indicate that PFF02 has good selectivity with an average z score of -3.46. Also the average RMSD for the lowest energy conformation for all these proteins is only 2.14 Å showing good predictability for a wide range of proteins. PFF02 thus emerges as a positive step towards a universal and transferable forcefield for all atom protein folding and tertiary structure prediction.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2,WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. Part of the simulations were performed at the KIST teraflop cluster and at the Barcelona Supercomputer Center.

References

1. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–6, 2001.
2. T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.*, 87:3100–9, 2004.
3. A. Verma and W. Wenzel. Towards a universal free-energy approach for all-atom protein folding and structure prediction. Submitted, 2007.
4. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181: 223–30, 1973.
5. A. Verma and W. Wenzel. Predictive and reproducible de-novo all-atom folding of a β-hairpin loop in an improved free energy forcefield. *J. Phys. Cond. Matt*, in press, 2007.
6. W. Wenzel. Predictive folding of a β-hairpin in an all-atom free-energy model. *Europhys. Lett.*, 76:156–162, 2006.
7. B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Molec. Biol.*, 258:367, 1996.
8. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker. Rosetta in CASP4: progress in ab-initio protein structure prediction. *Proteins*, 45:119–126, 2001.
9. J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 53:76–87, 2003.
10. A. Verma and W. Wenzel. Protein structure prediction by all-atom free-energy refinement. *BMC Structural Biology*, 7:12, 2007.